DOI 10.18699/SBB-2022-505

Detection of recombination events in coronavirus genomes from the subgenus sarbecovirus

Arefieva N.[1]*, Bukin Y.S.[1, 3], Dzhioev Y.P.[2], Zlobin V.I.[2]

[1] *Irkutsk State University, Irkutsk, Russia*

[2] *Limnological Institute, SB RAS, Irkutsk, Russia*

[3] *Irkutsk State Medical University, Irkutsk, Russia*

**arefieva.n4@gmail.com*

Key words: Recombination events, Sarbecoviruses, SARS-CoV-2

Motivation and Aim: The pandemic caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) poses a great threat to public health. Genetic recombination is one of the leading factors in the variability and evolution of viruses. Detection of recombination events in coronavirus genomes will provide new data on the possible origin and variability of SARS-CoV-2.

Methods and Algorithms: The complete genome sequences of the subgenus Sarbecovirus were downloaded from the NCBI GenBank database. The whole genomes of different lines of Sars-Cov-2 were downloaded from the GISAID database. The final sample contained 321 genomes. The annotation of open reading frames in genomes was performed using VADR (Viral Annotation DefineR) program. Multiple sequence alignment of complete genomes was performed in MAFFT v.7. The protein-coding genes 1ab, S, E, M, N present in all genomes present in all genomes were separated from whole-genome sequences. Each protein-coding gene was aligned according to the codon position using RevTrans 2.0. The aligned protein-coding parts were assembled into a fasta file. To detect potential recombination events in complete genomes, 9 algorithms included in RDP (Recombination Detection Program) v.5.5 software package were used (RDP, Genconv, Bootscan, Maxchi, Chimaera, SiSscan, PhylPro, LARD, 3Seq). The coordinates of the found recombination positions in the complete genomes were translated into coordinates for the coding part. The R script was written to verify the results and to separate recombination events from adaptive evolution and other evolutionary processes.

Results: Using different algorithms, 239 potential recombination sites were found. Results with insignificant p-value obtained by more than three algorithms were discarded. As a result, the number of recombination events decreased to 177. Most of the potential recombination events are located in the spike (S) protein gene, which is more susceptible to variability than other structural genes. The recombination events were detected between SARS-CoV-2, SAR-Cov-1, Bat coronavirus and Pangolin coronavirus strains.